

**CS989 research assessment**

**The statistics analysis**

**of National Football League**

**Quarterbacks**

**-Based on rookie quarterbacks of 2004-**

Word count

2895

**University of Strathclyde, Glasgow**

Table of Contents

[1.Introduction 1](#_Toc118207844)

[1.1 Dataset 1](#_Toc118207845)

[1.2 Research Objectives & Aims 2](#_Toc118207846)

[1.3 Methodology 2](#_Toc118207847)

[2. Analysis 2](#_Toc118207848)

[2.1 Data processing 2](#_Toc118207849)

[2.2 Passing yard comparison 4](#_Toc118207850)

[2.2 Pass complete rate analysis 5](#_Toc118207851)

[2.3 Total Touchdown analysis 6](#_Toc118207852)

[2.4 Passer ratings analysis 7](#_Toc118207853)

[2.5 Clustering Analysis of Quarterbacks 8](#_Toc118207854)

[2.5 Supervised method 10](#_Toc118207855)

[3. Conclusion 12](#_Toc118207856)

[4. Recommendation 12](#_Toc118207857)

[Appendix 13](#_Toc118207858)

[Reference list 14](#_Toc118207859)

# 1.Introduction

Quarterbacks in American football is treated as the most important position in American football, they are called commanders of offensive players in American football, it is because they lead the strategies of offensive play and lead the teammates during the game. Teams in NFL take care about having a great quarterback for their teams. (Wolfson, Addona and Schmicker, 2011). However, it is hard to find the great quarterback from college American football teams for their professional teams. Although NFL teams yield 453 million dollars in average (Gough, 2018) and Reyes and Swartz (2021) insisted that NFL made the highest profit among professional sports league, the data analysis of football is less developed than baseball and basketball (Baumer and Zimbalist,2014). NFL, however, also begins to invest to the big data analysis and holds the Kaggle competition which is called “Big Data Bowl” to find the insight of football plays from the big amount of data. Data also can make players objectively; it can prevent people measure players just the career of making teams have many trophies and single statistics. People also can find what quarterbacks are underestimated compared to their ability compared to the result of teams of their own or could not show their whole ability due to the wrong scheme of teams about match or the mismatch of players in the current teams. Data also can predict the performance of players (Craig and Winchester,2021), then it can be economically beneficial for the business strategy of NFL teams.

This research will deal with about the explanation of dataset, processing, and major statistics analysis, also about the research of representative statistics and machine learning analysis about the quarterback performance.

# 1.1 Dataset

The dataset for this research is found in Kaggle. This dataset includes the statistical indexes of performance of NFL quarterbacks from 1994 to 2016. This data consists of 14 variables.

“**QB**” means the name of quarterbacks,

“**ATT**” means that attempts of quarterbacks to pass to the offensive player. “CMP” indicates the frequency of complete pass.

“**YDS**” means the passing yard during the whole American football game, and it is measured by adding the distance from the line of scrimmage to the point that receiver received the ball and the distance that receiver rushed from the point that received the ball to the place where the offence place is ended.

“**YPA**” is yards per attempt, and it indicates that gained yardage by a pass play of Quarterback.

“**TD**” represents touchdown that quarterback made in the football match. Touchdown is like the “goal” in soccer, so that it is important for quarterbacks to make touchdowns

“**INT**” is interception, and this means the frequency of defensive players intercepted the ball

“**LG**” indicates Long, and it means the yardage of the longest pass in the match made by quarterback.

“**Sack**” is the term that Quarterback get tackled by defensive players and it means that fail to pass the ball.

“**Loss**” is the loss of yard that quarterback made.

“**Game\_points**” is the index that points scored in the game.

“**Rate**” means the Quarterback Rating.

“**home\_away**” indicates that whether the game was home game or away game.

“**year**” means that the year of the game was held.

# 1.2 Research Objectives & Aims

This is the research about the analysis of the performance of Quarterbacks from 2004 to 2016 based on major statistics of quarterbacks. There will be a performance analysis of quarterbacks, then attempt to find quarterbacks who made the best performance during 2004 to 2016 and made the performance like the best performance however, being underestimated due to the lack of career record about trophies. The research about quarterback evaluation statistic, passer rating will be done, and do the research about the relationship of passer ratings and touchdowns.

# 1.3 Methodology

The research will be performed from the data of quarterbacks from 2004 to 2016, the original dataset is consisted of the statistics of quarterbacks from 1996 to 2016, however, for the analysis, it will be processed and made into the new data frame. The reason of this is because most of franchise quarterbacks of NFL get drafted in this year. For example, Eli Manning ,Ben Roethlisberger, and Philip Rivers ,were drafted in 2004, and those quarterbacks are treated as a potential star quarterbacks at that time (Attner, 2004), then it is possible to find the statistic of veteran quarterbacks such as Tom Brady and Peyton Manning who are active before 2004, and It is possible to find the statistic of rookie quarterbacks who got drafted from 2010, for example, Russell Wilson, Cam Newton. The data is based on quarterbacks of 2004 then quarterbacks who are active before 2004 will be a group of veteran quarterbacks then quarterbacks who were active after 2010 will be treated as the rookie quarterback group . Analysis will be done by making an analysis of top 25 quarterbacks on each of statistics then unsupervised and supervised method will be done.

# 2. Analysis

# 2.1 Data processing

The original dataset contains the statistics of quarterbacks from 1996 to 2016, however, the research aim is analysis of dataset from 2004 to 2016 of single statistics and total statistics during the career of each quarterback. So, the first process of analysis is processing the original- dataset to the single game statistics of quarterbacks from 2004 and the total dataset of quarterbacks from 2004.

Table

Description automatically generated

Table

Description automatically generated

Fig 1. The original single match statistics data frame and data frame after processed

After data is processed, for the total statistics of quarterback, dataset of total statistic of quarterback is made like below.

Graphical user interface, table

Description automatically generated with medium confidence

Fig 2. The dataset of Total statistics of quarterbacks

Now dataset is processed, the dataset of total statistics for the total performance analysis, and single match performance will be used for the unsupervised and supervised method. Statistics “lg”, “game\_points”, “home\_away”, “int” were removed, it is because “game\_points” and “home\_away” are not related to quarterbacks, and lg(long pass) and interception may possibly be a good factor for evaluation but more important thing is to make the complete pass and making a big yard, so that these factors are removed.

# 2.2 Passing yard comparison

First is analysis of Quarterbacks about the passing yard, it is because Passing yard is one of the factors to analyse the performance of quarterback and Boulier et.al. (2010) also insisted that passing yard can be the good factor to evaluate quarterbacks , and it shows that how quarterback can gain the yard in the football game. From the Fig4, it also shows that passing yard and touch down have a positive relationship,

Chart, scatter chart

Description automatically generated

Fig 3. Relationship between passing yards and Touchdown

For the analysis, bar charts of top 25 quarterbacks. The reason of choosing top 25 is that It is possible to see the data of veterans and rookie quarterbacks.

Chart, bar chart

Description automatically generated

<Fig.4 Bar chart of passing yard of Top25 quarterbacks>

The graph above shows the top25 ranking of total passing yards of quarterbacks since 2004 to 2016, From this chart, Drafted quarterbacks in 2004 which are Eli Manning, Ben Roethlisberger and Philip Rivers had shown their good performance based on passing yards. Other Quarterbacks who got drafted later than 2004 also shows the good performance based on total passing yards. Andy Dalton, Andrew Luck, Came Newton, Ryan Tannehill, and Russell Wilson can be an example, based on this plot, these quarterbacks have a potential that can be the elite quarterback. From this chart It shows that two quarterbacks, Eli Manning and Ben Roethlisberger who got drafted in 2004 shows that they deserve to get their two Super bowl champion titles. It also shows that Philip rivers, even though he could not win his Super bowl champion title during his career, but this passing yard and his passing ability shows that he is also has great passing ability.

# 2.2 Pass complete rate analysis

Pass complete rate shows that each quarterbacks success rate of the pass. From this data the idea that the pass complete rate of quarterbacks can indicate the ability of passing skills of quarterbacks was gained, so that the bar chart of top 25 quarterbacks is made like below. From the Bar plot, it is possible to find that Drew Brees showed the highest rate about success of pass then from the chart of passing yard Drew Brees also showed the highest passing yards. From the chart It can be found that Philip Rivers is the seventh higher rate of pass success and Ben Roethlisberger has the nineth higher pass complete rate. From this chart, the opinion about Philip Rivers can be underestimated quarterback with great ability as a quarterback may possibly be strengthen. This also shows that Eli Manning had a fourth less rate about pass complete, however, he still had more than 60% of pass complete rate, also It shows that not only quarterback but also scheme of teams about the offense system can be important.

Background pattern

Description automatically generated with low confidence

<Fig.5 Bar chart of pass complete rate (pass success rate) of Top25 quarterbacks>

From this plot, it also shows that Rookie Quarterbacks who drafted after 2010 show that they also have similar pass complete rate to veteran Quarterbacks. Rookies such as Andy Dalton, Andrew Luck, Came Newton, Ryan Tannehill, and Russell Wilson and so on, showed that their passing yards can be explained with this pass complete rate. So, this stat can be the supplementary stat of the Passing Yards of Quarterbacks.

# 2.3 Total Touchdown analysis

The next investigation is about the Touchdown that quarterbacks made from 2004 to 2016. Because American football is not a sport to just make a lot of passes but must make a score to win the match. So, the ability to make touch down is necessary for quarterbacks. For this reason, research about total touchdown is done. For the research about touchdown the bar chart is plotted like below.

Histogram

Description automatically generated

<Fig.6 Top 25 Total touchdown of NFL quarterbacks >

The chart above shows the Top 25 ranking of total touchdown of quarterbacks, and This data also shows that even though veteran quarterbacks were active earlier than 2004 but still it shows that their performance to make a touchdown was great, It shows that Drew Brees made the highest touchdown record from 2004, although he drafted 1 year later than Tom Brady, He shows higher touchdown record than Tom Brady during the period of 2004 and 2016. Draftees in 2004 also shows the impressive performance, Eli Manning, even though he had gotten underestimated by people that not as better as his brother Peyton Manning, but in the perspective of making touchdown, he achieved to the fourth highest total touchdown record, and it is the highest record among draftees of 2004. From this, Eli Manning and Ben Roethlisberger showed that they contributed to their teams becoming Superbowl champion twice while they were playing for their teams. It also shows that rookie quarterbacks like Cam Newton and Russell Wilson, Andrew Luck also have the potential to be a good quarterback as a good passer. It also shows that other quarterbacks such as Jay Cutler shows the distinctive performance but can be underestimated due to the lack of Super bowl rings or record of football teams about going playoff or wildcard game.

# 2.4 Passer ratings analysis

Passer Rating is major stat to measure the performance of quarterbacks, this stat is made with complicated formula and it starting from minimum rate of 0 to the maximum rate of 158.3 (Reyers and Swartz, 2021). According to Dohlen(2011), the formular of Passer rating is like below.

Graphical user interface, application, table

Description automatically generated

Fig.7 The formula of Passer rating. Source: Journal of Quantitative Analysis in Sports

Due to the various aspects of the passing play is related, Craig and Winchester (2021) insisted that this Quarterback Rating is highly related to the win percentage of Quarterbacks. Heatmap (Fig10below shows correlations of each statistics of quarterbacks. From the heatmap, It shows that “rate” does not explain the pass complete or gained yard clearly, however, it shows the correlation more than 0.5 for the Yard per Attempt (YPA) and Touchdown(td) also has a 0.52 of correlation with the “game\_points”. From this perspective, Passer ratings can be seen that It explains the

Performance of quarterbacks to get the point. Passer Ratings, however, get criticised by people that it is biased for the specific game plan for modernised quarterbacks. Dohlen(2011) insists that main reason is because of Yard Per Attempt in formula is in the inverse proportion to the completions, for this reason It is required to modify the formula or making a new statistic.

Graphical user interface, chart, application, treemap chart

Description automatically generated

Fig8. Relationship between passing yards and Touchdown

According to Anderson (2016), ESPN already argued that Passer ratings has shortcomings due to this reason ESPN made a stat which is called Quarterback Ratings (which is called QBR). However, ESPN does not release the specific formula of QBR and only just releases their QBR on their homepage.

# 2.5 Clustering Analysis of Quarterbacks

The following analysis is about the clustering analysis of quarterbacks, and this is the research for quarterback grouping about the touchdown and K mean clustering was done

Shape

Description automatically generated

Fig9. Elbow test result for clustering

From the elbow method it shows that when it has three clusters, it shows the proper amount of clustering analysis. The graph indicates that 3 clusters will be an appropriate cluster amount.

A screenshot of a computer

Description automatically generated with low confidence

Fig.10 Completeness score and Homogeneity score result

Chart, scatter chart

Description automatically generated

Fig 11. K-Means clustering result

Fig11 and 12 indicate the Clustering test result and the cluster of Quarterbacks, then from this chart about touchdown, quarterback can be divided to the three types of playing styles of quarterbacks making touch down.

# 2.5 Supervised method

Logistic regression, there will be an investigation about touchdown prediction. Method had been done with the same data when clustering was used. The reason is to find the single game performance prediction, and single game statistic data has more data than combined statistic dataset. But also, statistics which are not related to the research are removed.

Text

Description automatically generated

Fig12. Correlation

The correlation shows that Passer rating has the highest relationship, so the study has been done with the “Touch Down” and “Passer Rating”.

Chart, bar chart

Description automatically generated

Fig 13. Barplot of Touchdown and Passer rating

Bar chaText

Description automatically generatedrt shows the relationship between Passer rating and Touchdown of quarterbacks. The data shows that quarterbacks can make the touchdown from 0 to 7 of it, then making more than 4 touchdowns can be treated a good quality of quarterback. So the data was processed and binarized, and dataset was trained for logistic regression model.



Fig 14. Training data shape and accuracy from model training

From the training of data, it shows the shape like “Fig14”, and it shows the 96.75% of the accuracy in the model. It can be plotted as a confusion matrix like below

Chart, treemap chart

Description automatically generated

Fig 15. Confusion Matrix

The confusion matrix shows that quarterbacks can make a touchdown TN has 2311 then FP has 13 also FN is 66, the value of TOP is 38. It shows that model has a high accuracy on predicting an ability of quarterbacks to make touchdown based on passer ratings. From this, although passer rating is criticised, passer rating still has great performance on the single game performance, and this is the reason that passer rating can be the major statistic of evaluating quarterback.

# 3. Conclusion

This is the research about investigating statistics of quarterbacks during from 2004 to 2016 and finding a player who showed the best performance during this period in the perspective of various aspect. Then research shows that via data of quarterbacks, it was possible to find the players who showed the best performance and what players has shown similar performance. It also can be possible to find the underestimated players. The method to quantify performance of quarterbacks by inventing new statistics for players, such as Passer rating can be a good way to measure quarterback performance objectively could be found during the research. However, statistical index of players also may possibly be the error and cannot be always correct.

In this perspective, it is important to make additional research about statistics of performance of quarterbacks. It is also necessary to do the research about the statistics of other positions, for example wide receivers, running backs, and offensive lineman, it is because each position in the offense is working together. From the case of ESPN QBR and big data bowl, which is competition held by NFL, the prospect of American football data analysis will be seen optimistic.

# 4. Recommendation

From this research, there was an opportunity to discover about the analysis method of American football and various method. It can be found that there could be the various method about collecting data from the football training and matches. There also are various analysis method to estimate players, and this can affect the various area, however, dataset only included essential statistics of quarterbacks so that there can be the limitation about doing a detailed analysis on each statistic. American football also is not the sports that one quarterback can carry the whole match, so that it is necessary to analyse quarterback then have a proper scheme of the team and ability of other offensive players. So, as it is mentioned, it is important to do the research about statistics of another position, then also about the analysis of match strategies and game plan too. Rushing yard of quarterback statistic also can be important to make the dataset, it is because this is because this statistic will make a correct group analysis of quarterbacks.

# Appendix

Edi

Python version

3.9.12 (main, Apr 5 2022, 01:53:17)

Jupyter Notebook version

6.4.8

Library for analysis :

* Pandas
* seaborn
* matplotlib
* numpy
* sci-kit learn
  + from sklearn, datasets
  + from sklearn , metrics
  + from sklearn import cluster
  + from sklearn.metrics import silhouette\_score
  + from sklearn.preprocessing , scale
  + from sklearn.metrics ,silhouette\_score
  + from sklearn.metrics ,homogeneity\_score
  + from sklearn.metrics ,completeness\_score
  + from sklearn.decomposition ,PCA
  + from sklearn.cluster, KMeans
  + from matplotlib, style
  + warnings
  + warnings.filterwarnings('ignore')
  + from sklearn.model\_selection train\_test\_split
  + from sklearn.metrics 🡪 confusion\_matrix, classification\_report, accuracy\_score, ConfusionMatrixDisplay
  + from sklearn.preprocessing StandardScaler
  + from sklearn.linear\_model LogisticRegression
  + from sklearn.tree DecisionTreeClassifier
  + from sklearn.ensemble RandomForestClassifier

# Reference list

Reyers, M. and Swartz, T.B., 2021. Quarterback evaluation in the national football league using tracking data. *AStA Advances in Statistical Analysis*, pp.1-16.

Gough, C. (2018). Topic: National Football League (NFL). [online] Available at: https://www.statista.com/topics/963/national-football-league/#dossierKeyfigures.

Birnbaum, J. (2021). Major Sports Leagues Lost Jaw-Dropping Amount Of Money In 2020. [online] Forbes. Available at: https://www.forbes.com/sites/justinbirnbaum/2021/03/06/major-sports-leagues-lost-jaw-dropping-amount-of-money-in-2020/?sh=4783b92a69c2 [Accessed 19 Oct. 2022].

Baumer, B. and Zimbalist, A.S. 2014. The sabermetric revolution : assessing the growth of analytics in baseball. Philadelphia: University Of Pennsylvania Press.

Wolfson, J., Addona, V. and Schmicker, R.H., 2011. The quarterback prediction problem: Forecasting the performance of college quarterbacks selected in the NFL draft. Journal of Quantitative Analysis in Sports, 7(3).

Attner, P., 2004. Above a cut: Steelers rookie Ben Roethlisberger is the gem among a group of four young quarterbacks who are providing an influx of refreshing new talent at the NFL's marquee position. The Sporting News, 228(47), pp.16-21.

Craig, J.D. and Winchester, N., 2021. Predicting the national football league potential of college quarterbacks. European Journal of Operational Research, 295(2), pp.733-743.

Volz, B.D., 2017. Race and quarterback survival in the National Football League. Journal of Sports Economics, 18(8), pp.850-866.

Boulier, B.L., Stekler, H.O., Coburn, J. and Rankins, T., 2010. Evaluating national football league draft choices: The passing game. *International Journal of Forecasting*, *26*(3), pp.589-605.

Smith, J., 2019, From Stacking to Tasking: Throwing, Running, and Race in Quarterbacks. Pacific University of Oregon Research repository[online], Available at :

<https://commons.pacificu.edu/work/sc/57caeaf0-c31f-43b7-ad5a-41a503661fbb> [Accessed : 01 NOV, 2022]

Data Source

LITTIEBRANT, J.2016. Quarterback Stats from 1996 – 2016.[online]

Available at : https://www.kaggle.com/datasets/speckledpingu/nfl-qb-stats

[Accessed : 20 OCT, 2022]